Group Project Milestone 2:

Analytics Implementation - Model Building & Evaluation

Group 5 Members:

Fajar Ibnu Fatihan

Li Jiayi

Lim Ming Wei Jerald

Wang Jinyue

National University of Singapore

IT5006 Fundamentals Of Data Analytics

AY2025/2026 Semester 2

1. Problem Definition & Data Preparation

1.1. Problem Definition

This project aims to develop two machine learning models to predict the 30-day readmission rate and the length of stay in the hospital of diabetic patients. These two metrics are key metrics of healthcare operational efficiency, clinical outcomes and systemic cost (Liu et al., 2024). The ability to predict the 30-day readmission of a patient would help identify high-risk patients during their initial hospital stay. This enables healthcare providers to implement targeted interventions in a timely manner to reduce the likelihood of readmission. The ability to predict a patient's length of stay would also help hospitals to better plan their resources around the needs of the patients.

1.2. Data Preparation and Feature Engineering

The initial dataset "Diabetes 130-US Hospitals for Years 1999-2008" provided by the University of California Irvine Machine Learning Repository contained 101,766 patient encounters and 50 features. A comprehensive data preparation pipeline was implemented to create a clean, high-quality dataset suitable for machine learning applications. This process involved several key stages, executed iteratively across our analysis notebooks.

1.2.1. Initial Data Cleaning and Reduction

In the dataset, five features had a large number of missing values and were removed. These include weight (96.9%), max_glu_serum (94.7%), A1Cresult (83.3%), medical_specialty (49.1%), and payer_code (39.6%). The missing values of the 'race' column were imputed with 'Other'.

It was also noticed that some of the features or records lend little predictive power and were removed. These include the features 'examide' and 'citoglipton', which had only one unique value, as well as the identifier columns of 'encounter_id' and 'patient_id'. Three records that had the 'gender' value unknown were also removed since the three records are insignificant to the total number of records.

The next set of records removed was deemed to potentially introduce bias to the prediction models. Where patients had already been readmitted previously, these records with duplicate 'patient_id' were removed while retaining only the first instance of the patient having been admitted. Records that indicate that patients were transferred to a hospice or expired were also removed given that these patients were either highly unlikely or unable to be readmitted to the hospital.

1.2.2. Feature Engineering and Transformation

Complex categorical features with many unique values were simplified by reducing the grouping of closely related values into an umbrella value. The age feature, originally in bins grouped by every 10 years, was simplified into broader categories of 'Young (<30)', 'Middle (30-49)', 'Older (50-69)', and 'Elderly (70+)'. Similarly, the features ‘diag_1’, ‘diag_2’, ‘diag_3’ had values grouped into ten major disease groups (e.g. ‘Circulatory’, ‘Respiratory’, ‘Diabetes’, etc.) according to ICD-9 diagnosis codes for simplified analysis and to reduce feature sparsity. For the columns ‘admission_type_id’, ‘discharge_disposition_id’ and ‘admission_source_id’, values were grouped into logical bins based on their provided descriptions. Their original feature columns were then dropped.

Several ratio features were engineered to capture the intensity of care relative to the severity of the case. These include ‘num_meds_per_diagnosis’, ‘procedures_per_diagnosis’, and ‘labs_per_diagnosis’. A ‘care_intensity_score’ feature was introduced as a weighted sum of key variables as these 4 key variables had the highest predictive power. A composite feature, adm_dis_path (e.g., ‘Emergency_to_Home’), was created, and binary variables were generated for the 10 most common pathways.

Encoding and normalisation were applied at the end of the feature engineering to ensure that the categorical features were compatible with the different models to be employed. Numerical features were normalised using StandardScaler to prevent any numerical feature from having an outsized effect on the learning models. For non-tree-based models, nominal categorical features were one-hot encoded into binary feature columns to prevent models from assuming an ordinal relationship between values. The "drop='first'" strategy was used to mitigate multicollinearity.

In the 30-day readmission dataset, an additional log transformation was applied to the features ‘'number_outpatient', 'number_emergency', 'number_inpatient', 'total_visits', 'medication_intensity' to reduce skewness and kurtosis. This ensured that values from these features more resembled a normal distribution.

Lastly, the dataset was split using an 80:20 ratio to obtain a training dataset and test dataset respectively.

2. Model Implementation & Training

The two problems of predicting length of stay in the hospital and 30-day readmission have different natures. Given that the former is a regression task and the latter a binary classification task, their approaches deviate and will be explained separately. A five-fold cross validation was used in all model training pipelines.

2.1 Length of Stay

Feature Scaling

RobustScaler was applied to the numerical features in both the training and test sets. RobustScaler was chosen over a standard scaler to minimize the influence of potential outliers in the data.

Dimensionality Reduction

Principal Component Analysis (PCA) was performed on the scaled training data for use in the linear regression models. PCA was used to determine the number of components required to explain at least 95% of the total variance. It was found that 4 principal components could explain 95.21% of the total variance.The training and test data for the linear models were then transformed into this 4-component PCA space.

Model 1: Ordinary Least Squares (OLS) Regression

A standard LinearRegression model was trained on the dataset after PCA.

Model 2: Ridge Regression (L2 Regularisation)

Hyperparameter tuning via GridSearchCV was performed to find the optimal regularisation parameter α. The best α was found to be 100. The final model was trained with this optimal parameter.

Model 3: Lasso Regression (L1 Regularisation)

Hyperparameter tuning via GridSearchCV was performed to find the optimal regularisation parameter α. The best α was found to be 0.001. The final model was trained with this optimal parameter.

Model 4: Random Forest Classifier

A powerful, non-linear ensemble model was chosen and trained on the full 140-feature scaled dataset (without PCA). Key hyperparameter settings included n_estimators=400, max_depth=25, min_samples_split=5, and max_samples=0.8.

Model 5: Gradient Boosting Classifier

Another non-linear ensemble model was trained on the scaled dataset. Hyperparameter settings included n_estimators=400, a slow learning_rate=0.05, and the Huber loss function for robustness to outliers. Early stopping was also configured to prevent overfitting.

Stacking Ensemble

This advanced ensemble model was trained on the full, scaled feature set, combining the predictions of three diverse base estimators: Random Forest Classifier, Gradient Boosting Classifier, and XGBoost Classifier. The predictions from these base models were then fed into a Ridge Regressor which acted as the final meta-learner.

2.2 30-Day Readmission

SMOTE

A significant class imbalance was identified in the training data, where only 8.9% of records belong to the “readmitted” class. To reduce the class imbalance in the testing, two versions of synthetic minority oversampling technique (SMOTE) were used on the training dataset to balance the classes. The first being a classic SMOTE and the second SMOTEENN, or SMOTE followed by a cleaning process known as Editing Nearest Neighbours (EEN) to reduce noise in the dataset. With this added cleaning, SMOTEENN aims to reduce overfitting in the model and improve its ability to generalise.

Model 1: Logistic Regression

Hyperparameter tuning via GridSearchCV was performed to find the optimal regularisation parameter C. The best C was found to be 0.01 in the dataset with SMOTE and 1 in the dataset with SMOTEENN. The final model was trained with this optimal parameter.

Model 2: Random Forest Regressor

A powerful, non-linear ensemble model was chosen and trained on the full 44-feature dataset. Key hyperparameter settings included n_estimators=100, max_depth=10, min_samples_leaf=2 for the dataset with SMOTE and n_estimators=100, max_depth=20, min_samples_leaf=4 for the dataset with SMOTEENN.

Model 3: XGBoost Regressor

Extreme Gradient Boosting (XGBoost) is another powerful ensemble learning model. The best set of hyperparameters found using the SMOTE dataset were learning_rate = 0.1, max_depth = 10, n_estimators = 200 and subsample = 1.0. The best set of hyperparameters found using the SMOTEENN dataset were learning_rate = 0.05, max_depth = 5, n_estimators = 100 and subsample = 0.7.

Model 4: LightGBM

Light Gradient Boosting Machine (LightGBM) is an alternative to the other tree-based ensemble learning methods. It uses histogram-based decision tree learning and leaf-wise tree growth, and optimises speed and memory efficiency. From the GridSearchCV, the best set of hyperparameters found using the SMOTE dataset were learning_rate = 0.1, max_depth = 20, n_estimators = 200 and num_leaves = 50. The same for the SMOTEENN dataset were learning_rate = 0.05, max_depth = 20, n_estimators = 100 and subsample = 50.

3. Model Evaluation & Comparison

3.1 Length of Stay

Model performance was evaluated using three key metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Table 1 below is a summary of the metrics results from different models.

Model	Training RMSE (days)	Test RMSE (days)	Test MAE (days)	Training R^2	Test R^2
OLS Regression	2.6745	2.6330	2.0195	0.2019	0.2035
Ridge Regression (α=100)	2.6745	2.6330	2.0195	0.2019	0.2035
Lasso Regression (α=0.001)	2.6745	2.6330	2.0196	0.2019	0.2035
Random Forest Regressor	2.0670	2.3103	1.7508	0.5233	0.3868
Gradient Boosting Regressor	2.2653	2.2863	1.6971	0.4274	0.3994
Stacking Regressor	2.1182	2.2579	1.7002	0.4994	0.4143

Table 1. Summary of Metrics From Length of Stay Prediction Models

The performance of all three linear models, OLS, Ridge Regression, and Lasso Regression were nearly identical. The linear models achieved a Test R^2 of approximately 0.2035, indicating that they collectively explain only about 20% of the variance in the Length of Stay. The near-identical training and test RMSE values suggest the baseline linear model had minimal overfitting. Furthermore, the lack of improvement from Ridge Regression and Lasso Regression regularisation confirms that multicollinearity was not a significant issue after applying PCA.

Both non-linear ensemble models significantly outperformed the linear models. The Random Forest model showed a notable gap between its Training R^2 (0.5233) and Test R^2 (0.3868), suggesting some degree of overfitting to the training data, despite the use of regularisation techniques.

The Test RMSE was 2.2863 days from Gradient Boosting, which is an improvement over the linear models' 2.6330 days. The best Test MAE was 1.6971 days from Gradient Boosting, meaning its predictions are off by about 1.7 days on average.

The Stacking Regressor is the best-performing model, achieving the highest predictive power with a Test R^2 of 0.4143. This model explains over 41% of the variance, confirming that combining diverse models (RF, GB, XGB) effectively captures more complex patterns than any single model alone. The Stacking Regressor also achieved the lowest Test RMSE of 2.2579 days.

From Figure 1 in the appendix, it was noted that the top five features in the prediction of the length of hospital stay were the care intensity score, number of medications prescribed, number of lab procedures carried out, number of medications per diagnosis and the procedure intensity.

3.2 30-Day Readmission

The 30-day Readmission prediction models were evaluated based on their precision, recall, accuracy and weighted F1-scores summarised in Table 2.

Model	SMOTE Applied	accuracy	Class 0 (not readmitted)		Class 1 (readmitted)		weighted F1
			recall	precision	recall	precision	
Logistic Regression	SMOTE	0.6420	0.6501	0.9376	0.5592	0.1356	0.7187
	SMOTEENN	0.3106	0.2555	0.9532	0.8721	0.1031	0.3835
Random Forest	SMOTE	0.8922	0.9743	0.9131	0.0556	0.1750	0.8660
	SMOTEENN	0.7882	0.8331	0.9270	0.3312	0.1630	0.8186
XGBoost	SMOTE	0.9103	0.9990	0.9111	0.0064	0.3810	0.8689
	SMOTEENN	0.8120	0.8641	0.9244	0.2804	0.1684	0.8322
LightGBM	SMOTE	0.9104	0.9992	0.9110	0.0056	0.4118	0.8689
	SMOTEENN	0.8576	0.9229	0.9229	0.1922	0.1966	0.8569

Table 2. Summary of Metrics From 30-Day Readmission Prediction Models

The performance of the three non-linear models outperformed the Logistic Regression model. The performance of the Random Forest and XGBoost were also comparable with present literature by Liu et al. (2024), with F1 scores between 0.86 and 0.87. This indicates a good balance between precision and recall in these models' predictions of 30-day readmission of patients.

Of the three non-linear models, LightGBM narrowly edges out over XGBoost as the most performant model, with the highest accuracy score of 0.9104 and an F1 score of 0.8689. This balances the predictive power of the model in both the minority as well as the majority class.

It was also noted that the models trained on the SMOTEENN dataset consistently improved in their recall score while reducing their precision, accuracy and F1 scores. This indicates an enhanced ability to predict the minority class, at the expense of increased false positives. This could be due to the removal of ambiguous or borderline samples that would have otherwise been identified with the majority class.

From Figure 2 in the appendix, it was noted that the top five features in the prediction of 30-day readmission were the number of inpatient visits in the year prior to admission, whether the patient was discharged to home, total visits to the hospital in the year prior to admission, care intensity score and whether the patient was discharged to a rehab facility.

4. Model Interpretation & Business Insights

4.1 Length of Stay

Our analysis provides actionable insights into the factors that drive hospital length of stay. The engineered features like ‘care_intensity_score’, which combines the number of medications, procedures, labs, and diagnoses, and ‘procedure_intensity’, were highly predictive. This suggests that patients with longer stays are primarily those requiring a higher volume of medical services, not simply those with more diagnoses.

The importance of the highly ranked squared terms ‘num_medications_squared’ and ‘num_lab_procedures_squared’ implies that the relationship between the number of interventions and length of stay is non-linear and accelerating. This suggests that an increase in procedures and medications administered to a patient may indicate an exponential increase in length of stay.

It was also noted that the ratio features ‘num_meds_per_diagnosis’ and ‘medication_intensity’ contributed more to the models’ predictive power than just the raw number of diagnoses (Figure 1). This highlights that a longer stay is linked to the difficulty in treating the primary conditions, which can be measured by a higher number of medications or lab procedures relative to the number of diagnoses.

The feature ‘medical_specialty’ is the only highly-ranked categorical feature, which demonstrates that the type of physician and department managing the patient plays a role in length of stay, potentially reflecting differences in treatment protocols or patient complexity within those specialties.

4.2 30-Day Readmission

Our analysis, utilising a Random Forest model, revealed that a patient's history is the most critical predictor of readmission within 30 days of discharge. The number of prior inpatient visits stands out as the single most influential factor, which may explain the similarly highly ranked ‘total_visits’ which measures the overall frequency of inpatient, outpatient, and emergency visits. This would suggest that patients with a large number of visits to the hospital, especially inpatient visits, could be at higher risk of readmission and should be assessed more thoroughly before discharge.

Similar to the length of stay study, the calculated ‘care_intensity_score’ was also highly ranked as a predictor of a patient’s 30-day readmission. This taken together with the higher propensity for patients discharged to homes to be readmitted to the hospital within 30 days presents an alarming need to manage the discharge of these patients well. There is a possibility that the patient or primary caregiver is unable to care for patients at home, and should thus be avoided especially for patients who had gone through more intensive care as derived by the ‘care_intensity_score’. This could suggest the need for hospitals to recommend a transfer to a step-down care facility before allowing such patients to return home.

4.3 Applications of Prediction Models

Based on these findings, the primary recommendation is to embed these predictive models into the hospital's Electronic Health Record (EHR) system to proactively stratify patients by risk upon admission. This data-driven approach enables a shift from reactive to proactive care, allowing clinical resources to be targeted effectively. For high-risk individuals, this means initiating a more robust discharge process early in their stay, featuring detailed medication reconciliation by a pharmacist and pre-scheduled follow-up appointments. Furthermore, the model's output should trigger a "readmission risk huddle," where a multidisciplinary clinical team can collaborate to create a tailored care plan, thereby augmenting clinical judgment and improving patient outcomes.

5. Conclusion

In conclusion, this analysis successfully identifies the key drivers for both hospital length of stay and 30-day readmission, with patient complexity emerging as the central theme. For length of stay, the intensity of care is the most critical predictor. Similarly, for readmission risk, a patient's history, particularly the number of prior inpatient visits and overall healthcare utilisation, are the most influential factors.

Ultimately, the models demonstrate that a data-driven approach can effectively pinpoint high-risk patients. By integrating these predictive insights into the hospital's EHR system, the organization can shift from a reactive to a proactive care model. This allows for the strategic allocation of clinical resources, enabling targeted interventions like enhanced discharge planning and multidisciplinary team huddles for those most in need. Implementing these recommendations will lead to more efficient hospital operations, reduced readmissions, and significantly improved patient outcomes.

References

Liu, V., Sue, L., & Wu, Y. (2024). Comparison of machine learning models for predicting 30-day readmission rates for patients with diabetes. Journal Of Medical Artificial Intelligence, 7. doi:10.21037/jmai-24-70

Appendix

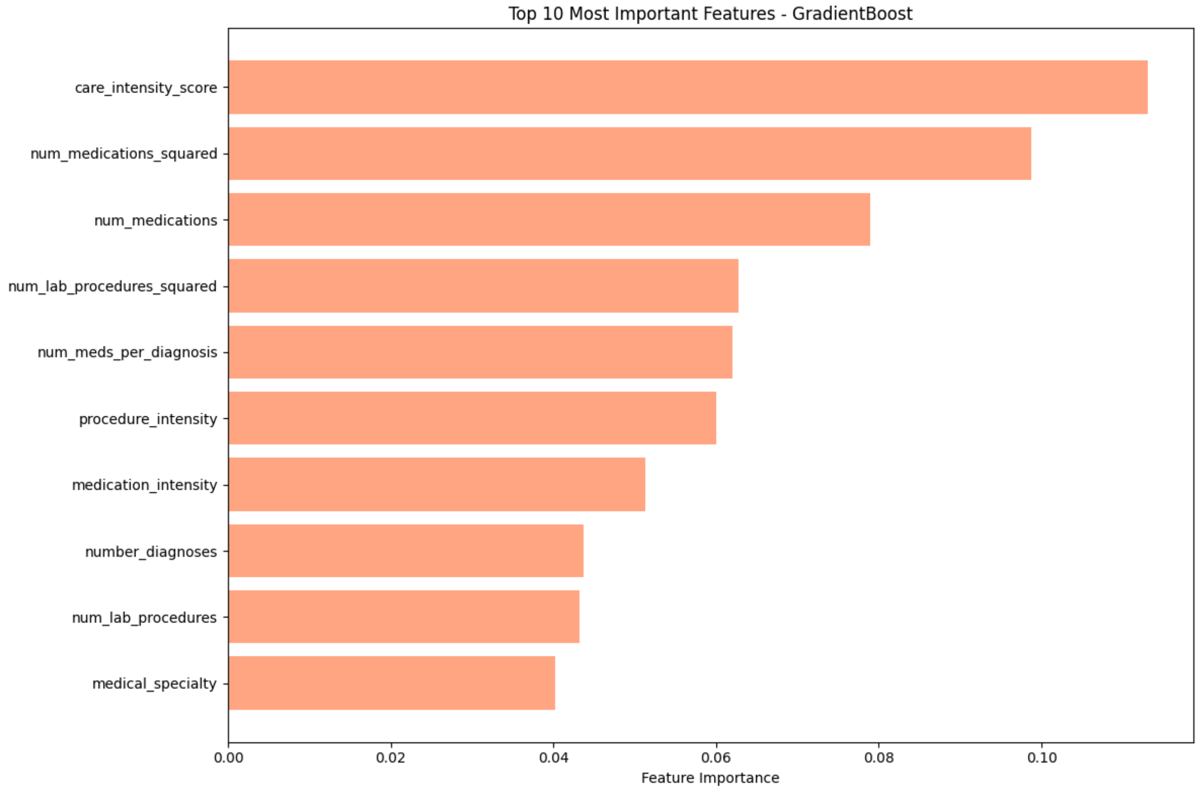

Figure 1. Top 10 Most Important Features - GradientBoost (Length of Stay)

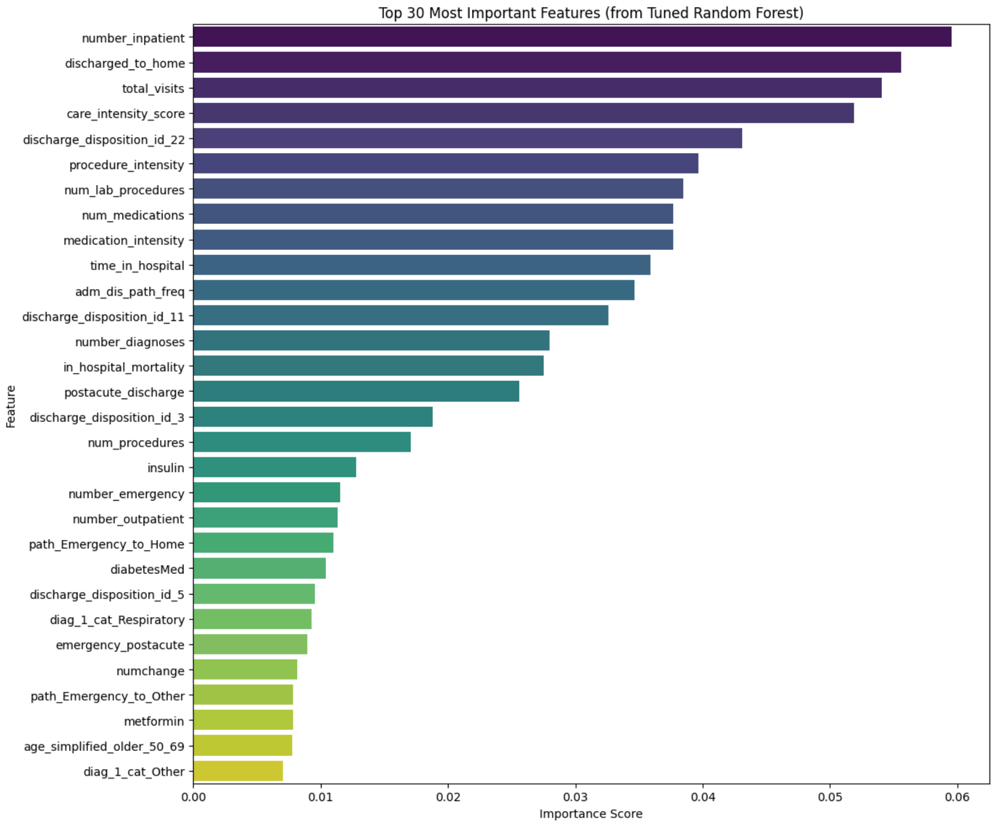

Figure 2. Top 30 Most Important Features - Random Forest (Readmissions)

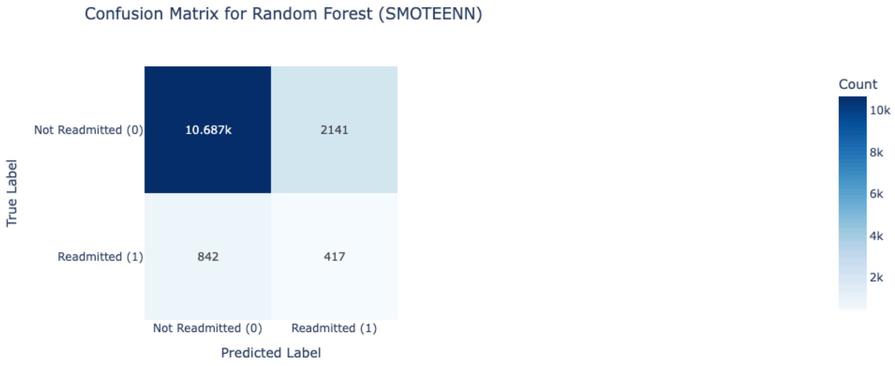

Figure 3. Confusion Matrix Iteration 4 - Random Forest - SMOTEENN (Readmissions)

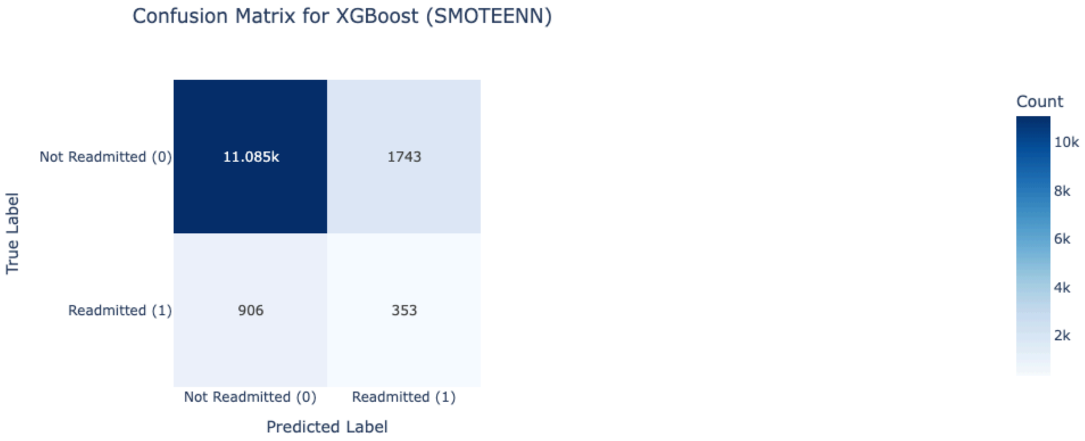

Figure 3. Confusion Matrix Iteration 4 - XGBoost - SMOTEENN (Readmissions)

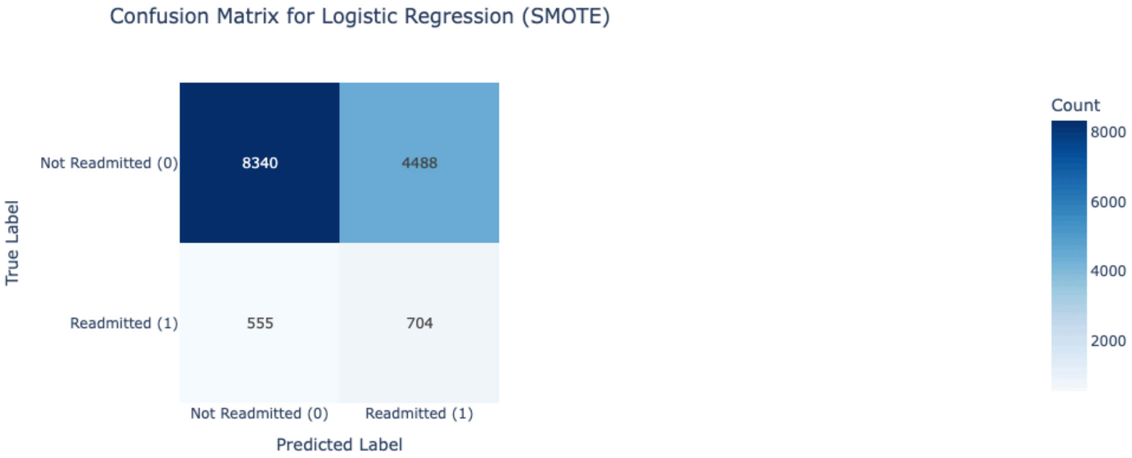

Figure 5. Confusion Matrix Iteration 4 - Logistic Regression - SMOTE (Readmissions)

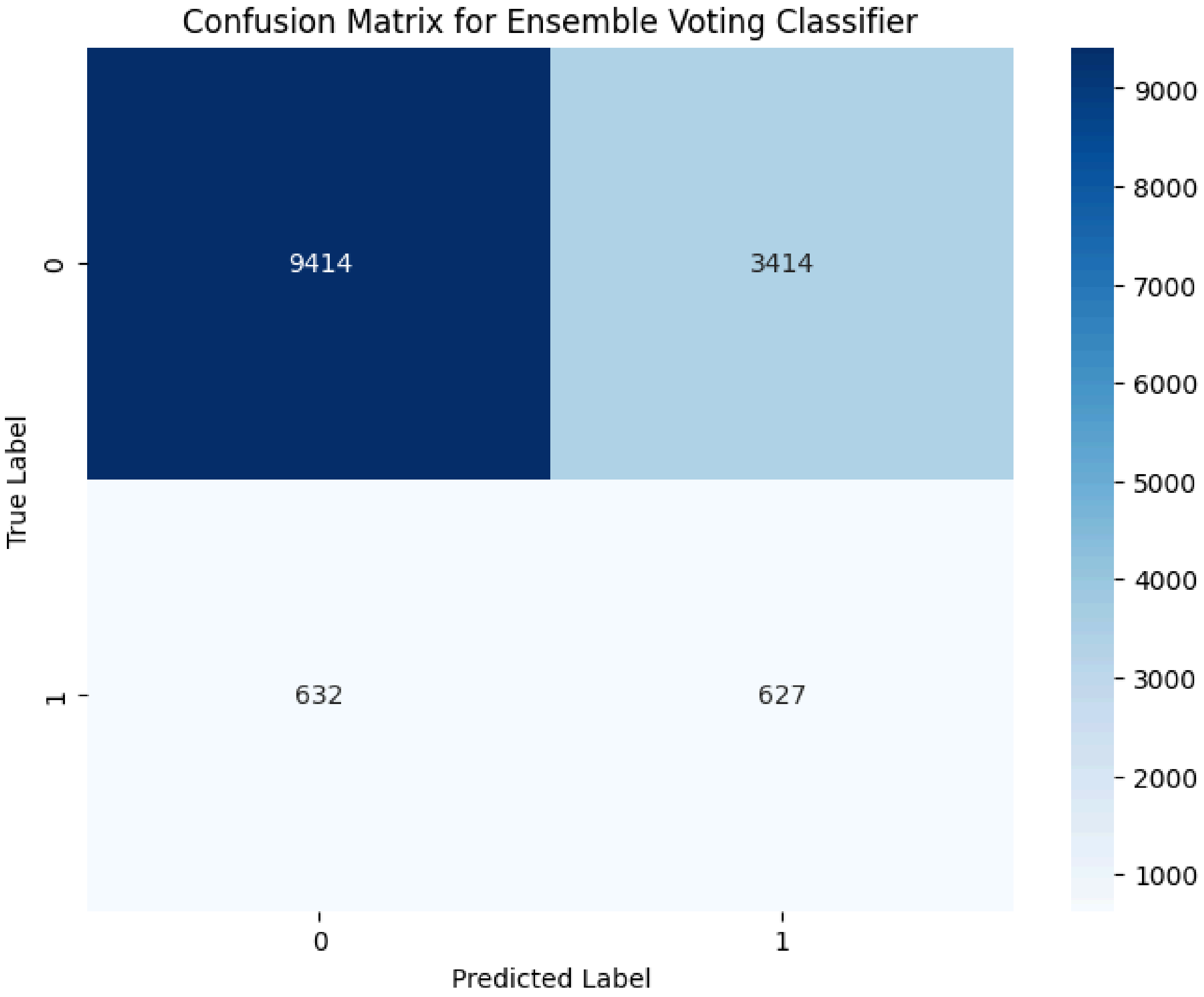

Figure 6. Confusion Matrix Iteration 5 - Ensemble Method - Voting Classifier (Readmissions)

AI Tool Declaration

We used GPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5 to generate and validate ideas, improve expression and refine our writing. We also used the models to debug our code. We are responsible for the content and quality of the submitted work.